

Fast calculation of two-electron-repulsion integrals: a numerical approach

Pedro E. M. Lopes

Rua Almirante Reis, N° 28A, 2° Esq, 2330-099 Entroncamento, Portugal *

An alternative methodology to evaluate two-electron-repulsion integrals based on numerical approximation is proposed. Computational chemistry has branched into two major fields with methodologies based on quantum mechanics and classical force fields. However, there are significant shadowy areas not covered by any of the available methods. Many relevant systems are often too big for traditional quantum chemical methods while being chemically too complex for classical force fields. Examples include systems in nanomedicine, studies of metalloproteins, etc. There is an urgent need to develop fast quantum chemical methods able to study large and complex systems. This work is a proof-of-concept on the numerical techniques required to develop accurate and computationally efficient algorithms for the fast calculation of electron-repulsion integrals, one of the most significant bottlenecks in the extension of quantum chemistry to large systems. All concepts and calculations were developed for the three-center integral $(p_{xA}p_{xB}|p_{xC}p_{xC})$ with all atoms being carbon. Starting with the analytical formulae, convenient decompositions were tested to provide smooth two-dimensional surfaces that were easily fitted. The approximating algorithm consisted of a multilayered approach based on multiple fittings of two-dimensional surfaces. An important aspect of the new method is its independence on the number of contracted Gaussian primitives. The basis set of choice was STO-6G. In future developments, larger basis sets will be developed. This work is part of a large effort aimed at improving the inadequacies of existing computational chemistry methods, both based on quantum mechanics and classical force fields, in particular in describing large and heterogeneous systems (ex. metalloproteins).

I. INTRODUCTION

The field of computational quantum chemistry has experienced extraordinary progress to date due to advances in computing power and the development of new algorithms. While advances have been reached, still there are limitations in the size and/or complexity of the systems that can be studied. In the second decade of the twenty-first century the words of Paul Dirac in 1929 [1] still echo: *“the underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble.”* Today, Dirac’s statement still remains true and many of the equations governing the chemical phenomena are still too complex to solve using today’s computational resources.

To answer complex chemical phenomena, computational quantum chemistry has suffered multiple numerical approximations and simplifications. Some are numerical approximations to the fundamental equations such as the Born-Oppenheimer approximation that greatly simplifies the Schrödinger equation by considering that the much heavier nuclei remain stationary during the calculation. Other approximations, leading for example to empirical and semi-empirical methods, consider simplified forms of the first-principles underlying equations that are typically faster to solve. Other classes of methods in computational chemistry have abandoned the quantum chemical principles altogether and classical approximations to the potential energy surface based on force fields were developed, as they are computationally less intensive than quantum chemical electronic structure calculations. Empirical force fields are currently the methods of

choice for studies of large systems in biology and materials science, for example conformational studies of proteins, DNA, etc and protein-ligand binding thermodynamics. However, empirical force fields have severe limitations: limited applicability, difficulty in describing complex chemistries and, inability to describe systems where formation and breakage of bonds occur. Empirical force fields are highly parameterized and typically include terms for bonds, angles and torsions plus non-bonding terms [2]. Empirical force fields are limited to the systems used in developing the parameters (ex. proteins, lipids, DNA/RNA, etc) and the parameterizations usually cover sp, sp² and sp³ hybridizations. It is extremely difficult to cover complex chemical spaces with force fields, for example when transition metals are involved. The harmonic nature of force fields does not typically allow for breaking and formation of chemical bonds. In contrast, high-level quantum chemical methods can describe most systems, but are still limited to small models, at least when compared to typical systems studied by classical force fields. QM/MM mix quantum chemical methods with empirical force fields and, thus, are able to describe large systems. QM/MM methods work better when the quantum region is highly localized but are inadequate to describe the dynamics of large systems.

Currently, many areas such as biophysics, biochemistry, materials science, nanomedicine, etc. cannot be described using existing methodologies. These systems have huge chemical spaces that are impossible to cover using existing empirical force field methods and are too big for current Quantum Mechanics (QM) techniques, even the best linear scaling methods. There is a clear “capabilities gap” in existing computational methodologies that need to be urgently addressed. Emerging fields such as nanomedicine or materials science would benefit from new computational methodologies based on QM. Traditional applications of classical force field methods would benefit as well. For example, it is estimated that half of all proteins are metalloproteins [3]. Simulations of

* 11 Warren Lodge Ct 1D, Cockeysville, MD 21030, USA; plopesuk@yahoo.co.uk ; www.fastcompchem.com

metalloproteins would greatly benefit from fast QM methods since existing classical force fields have problems describing such systems.

In summary, new methodologies are needed to bridge the “capabilities gap” between current quantum chemical methods and classical force fields. In the base Hartree-Fock method, the major contributors to the cost of the calculation are the computation of the two-Electron-Repulsion Integrals (ERIs), with a quartically scaling $O(N^4)$, diagonalization of the Fock matrix with a cubically scaling $O(N^3)$ and the self-consistent procedure that typically adds more than 10 iterations. Development of new computational methodologies based on QM will have to address each of the restrictions in order to achieve acceptable speeds. The aim of the current work is to develop an alternative technique, based on accurate numerical approximations, for the fast computation of ERIs. In computational quantum chemistry, the most common basis sets are based on Gaussian basis functions. It was already apparent in the 1950s that calculations of polyatomic systems based on Slater-type orbitals would be intractable. The breakthrough occurred when Boys proposed basis functions based on Cartesian Gaussian functions [4]. It was also found that linear combinations of Gaussians, designated as contracted Gaussians, could approximate atomic orbitals with great accuracy. Ever since, contracted Gaussians have been the basis set of choice, being used in all of the major program packages.

Computation of ERIs has a long history. Initially, all molecular integrals were calculated analytically since closed formulas for integrals over Gaussians were easily derived. The analytical formulas being specific to each integral do not allow the systematic calculation of integrals of higher angular momentum. Several recursive methodologies were then developed and gained acceptance in modern computational quantum chemistry programs. In this category are included the methods of Rys polynomials [5,6], McMurchie and Davidson [7] and Obara and Saika [8]. More recently, active work has been developed on approximate methodologies to speed up the computation of ERIs, for example, approaches using density fittings or the Cholesky decomposition. A very good and recent review of the calculation of ERIs has been published by Reine *et al.* [9].

The methodology to compute ERIs proposed in this work differs in concept and praxis relative to previous and current approaches. Existing methodologies need to be generic and applicable to any basis set. In contrast, the method being proposed approximates a pre-determined basis set and is optimized for speed, as it needs to be many orders of magnitude faster than current methods. The new computational methodology will use single- ζ (double- ζ for transition metals) basis sets. This paper is a proof-of-concept on the development of accurate numerical approximations to the analytical formulae for ERIs. The work will focus on the integral $(p_{xA}p_{xB}|p_{xC}p_{xC})$ with atoms A, B, and C being carbon. The choice of three-

center integrals offers significant advantages. The numerical approximations are simpler in three-center ERIs than in four-center ERIs because a smaller number of coordinates are required. By keeping all elements the same, significant symmetry relationships are introduced and smaller domains for the coordinates can be considered (see Sect. IV; Figure S1), thus reducing the number of target points that are required for the numerical approximations.

II. THEORETICAL BACKGROUND

A. Revisit the analytical calculation of two-electron-repulsion integrals

In the early years of computational chemistry, ERIs were calculated using analytical formulae [10]. The notation used for the explicit expressions of ERIs over Cartesian Gaussian functions is kept as close as possible to the one used by Clementi [11]. An important concept in molecular orbital theory is the expansion of the basis functions $\phi_i(A)$ as linear combinations of primitive Cartesian Gaussian Type functions (GTFs):

$$\phi_i(A) = \sum_{a=1}^n c_{ia} \eta_a(A) \quad (1)$$

Cartesian GTFs are composed of a radial Gaussian function multiplied by Cartesian coordinates x, y and z with exponents l_i, m_i , and n_i ,

$$\eta_i(A) = x_A^{l_i} y_A^{m_i} z_A^{n_i} \exp(-\alpha_i r_A^2) \quad (2)$$

The basic steps required to derive the ERI of the kind $(p_{xA}p_{xB}|p_{xC}p_{xC})$ are briefly described. Testing the novel numerical algorithms on three-center ERIs is important because they are significantly simpler than the four-center counterparts, due to having fewer degrees of spatial freedom, while still requiring the same techniques to perform the approximation. ERIs over basis functions are themselves written as linear combinations over the primitive GTFs:

$$(\phi_A \phi_B | \phi_C \phi_C) = \sum_{a,b,c,c} c_a c_b c_c c_c (\eta_A \eta_B | \eta_C \eta_C) \quad (3)$$

The advantage of using GTFs stems from the Gaussian product theorem is that the product of two GTFs is another GTF. In Eq. 3 the product of the first pair centered at \vec{A} and \vec{B} results in the general formula:

$$\eta(\alpha_1, \vec{A}, l_1) \eta(\alpha_2, \vec{B}, l_2) = \exp\left(-\frac{\alpha_1 \alpha_2 (\overline{AB})^2}{\gamma_1}\right) \times \sum_{i=0}^{l_1+l_2} f_i(l_1, l_2, \overline{PA}_x, \overline{PB}_x) x_P^i \exp(-\gamma_1 x_P^2) \times \sum_{j=0}^{m_1+m_2} f_j(m_1, m_2, \overline{PA}_y, \overline{PB}_y) y_P^j \exp(-\gamma_1 y_P^2) \times \sum_{k=0}^{n_1+n_2} f_k(k_1, k_2, \overline{PA}_z, \overline{PB}_z) z_P^k \exp(-\gamma_1 z_P^2) \quad (4)$$

with

$$\gamma_1 = \alpha_1 + \alpha_2 \quad (5)$$

and

$$\vec{P} = \frac{\alpha_1 \vec{A} + \alpha_2 \vec{B}}{\gamma_1} \quad (6)$$

Similar equations can be derived for the second pair with:

$$\gamma_2 = \alpha_3 + \alpha_4 \quad (7)$$

and

$$\vec{Q} = \frac{\alpha_3 \vec{C} + \alpha_4 \vec{D}}{\gamma_2} \quad (8)$$

The functions f_i, \dots, f_k appearing in Eq. 4 result from the application of the binomial theorem to the products of Gaussian functions. Their generic formula is:

$$f_i(l_1, l_2, A, B) = \sum_{j=\max(0, l_1-l_2)}^{\min(i, l_1)} \frac{l_1! l_2! A^{l_1-j} B^{l_1-j}}{j! (l_1-j)! (l_2-i+j)! (l_2-i+j)!} \quad (9)$$

Explicit values of the function $f_i(l_1, l_2, A, B)$ are given in Table I up to $l_1+l_2=4$. Substituting the pairs $\eta(\alpha_1, \vec{A}, l_1) \eta(\alpha_2, \vec{B}, l_2)$ and $\eta(\alpha_3, \vec{C}, l_3) \eta(\alpha_4, \vec{D}, l_4)$ into Eq. 3 results in the formal formula for the analytical calculation of ERIs (Eq. 10) where the normalization factors are written as N_α :

$$(\phi_A \phi_B | \phi_C \phi_D) = \sum_{a,b,c,c} c_a c_b c_c c_d (\eta_A \eta_B | \eta_C \eta_D) = \sum_{a,b,c,c} c_a c_b c_c c_d \exp\left(-\frac{\alpha_1 \alpha_2 (\overline{AB})^2}{\gamma_1}\right) \exp\left(-\frac{\alpha_3 \alpha_4 (\overline{CD})^2}{\gamma_2}\right) \times \sum_{i=0}^{l_1+l_2} f_i(l_1, l_2, \overline{PA}_x, \overline{PB}_x) \times \sum_{j=0}^{m_1+m_2} f_j(m_1, m_2, \overline{PA}_y, \overline{PB}_y) \times \sum_{k=0}^{n_1+n_2} f_k(k_1, k_2, \overline{PA}_z, \overline{PB}_z) \times \sum_{i'=0}^{l_3+l_4} f_{i'}(l_3, l_4, \overline{QC}_x, \overline{QD}_x) \times \sum_{j'=0}^{m_1+m_2} f_{j'}(m_3, m_3, \overline{QC}_y, \overline{QD}_y) \times \sum_{k'=0}^{n_3+n_4} f_{k'}(k_3, k_4, \overline{QC}_z, \overline{QD}_z) \iint x_P^i y_P^j z_P^k x_Q^{i'} y_Q^{j'} z_Q^{k'} \frac{1}{r_{12}} \exp(-\gamma_1 r_{P_1}^2 - \gamma_2 r_{Q_2}^2) dV_1 dV_2 \quad (10)$$

A simplified notation, $\left\{x_{P_1}^i y_{P_1}^j z_{P_1}^k | x_{Q_2}^{i'} y_{Q_2}^{j'} z_{Q_2}^{k'}\right\}$, is introduced for the integral $\iint x_P^i y_P^j z_P^k x_Q^{i'} y_Q^{j'} z_Q^{k'} \frac{1}{r_{12}} \exp(-\gamma_1 r_{P_1}^2 - \gamma_2 r_{Q_2}^2) dV_1 dV_2$ in the remaining of the text.

Calculation of ERIs according to Eq. 10 requires repeated evaluations of $\left\{x_{P_1}^i y_{P_1}^j z_{P_1}^k | x_{Q_2}^{i'} y_{Q_2}^{j'} z_{Q_2}^{k'}\right\}$, "f" functions, normalization factors and the two exponential functions, $\exp\left(-\frac{\alpha_1 \alpha_2 (\overline{AB})^2}{\gamma_1}\right)$ and $\exp\left(-\frac{\alpha_3 \alpha_4 (\overline{CD})^2}{\gamma_2}\right)$, over multiple

loops. There are loops over the contraction coefficients, c_a, c_b, c_c and c_d , and the indices i, j , etc. The indices i, j , etc. determine the f functions and the integrals $\left\{x_{P_1}^i y_{P_1}^j z_{P_1}^k | x_{Q_2}^{i'} y_{Q_2}^{j'} z_{Q_2}^{k'}\right\}$. The index i runs between 0 and $l_1 + l_2$ and similarly for j, k, i', \dots , which depend on $m_1 + m_2, n_1 + n_2, l_3 + l_4, \dots$. When

the exponents i, j, \dots are zero, the corresponding $x_{P_1}^i, y_{P_1}^j, \dots$ terms are indicated as "1" in $\left\{x_{P_1}^i y_{P_1}^j z_{P_1}^k | x_{Q_2}^i y_{Q_2}^j z_{Q_2}^k\right\}$.

Although the complexity of the integrals $\left\{x_{P_1}^i y_{P_1}^j z_{P_1}^k | x_{Q_2}^i y_{Q_2}^j z_{Q_2}^k\right\}$ increases with larger values of l_1, l_2, \dots , each is a well-defined function of \vec{P} and \vec{Q} through the distances \overline{PQ} and the corresponding non-zero projections along the Cartesian axis \overline{PQ}_x (also $\overline{PQ}_y, \overline{PQ}_z$, when the integral involve y- and z- functions). Recalling the definitions of \vec{P} and \vec{Q} from Eqs 6 and 8, respectively, the integrals $\left\{x_{P_1}^i y_{P_1}^j z_{P_1}^k | x_{Q_2}^i y_{Q_2}^j z_{Q_2}^k\right\}$ are functions of the coordinates of A, B, C, and D (A,B, and C in the three-center case).

The main purpose of this work is to illustrate how the three-center integrals $(p_{xA}p_{xB}|p_{xC}p_{xC})$, when carbon atoms are placed at the centers A, B, and C, can be calculated accurately through numerical approximation. In the following Sections, the specific simplifications introduced by the considering a three-center ERI, and the mathematical details of the numerical approximations are discussed.

B. Numerical fitting of three-center two-electron-repulsion integrals ($\mathbf{p}_{xA}\mathbf{p}_{xB}|\mathbf{p}_{xC}\mathbf{p}_{xC}$)

Calculation of three-center ERI $(p_{xA}p_{xB}|p_{xC}p_{xC})$ involves significant simplifications resulting from many of the "f" functions becoming null, according to Table I. After the null terms are omitted, Eq. 10 can be rewritten as:

$$(p_{xA}p_{xB}|p_{xC}p_{xC}) = \sum_{a,b,c} c_a c_b c_c^2 N_a N_b N_c^2 \exp\left(-\frac{\alpha_1 \alpha_2 (\overline{AB})^2}{\gamma_1}\right) \times \\ \left[\overline{PA}_x \overline{PB}_x \{111|x_{Q_2}^2 11\} + (\overline{PA}_x + \overline{PB}_x) \{x_{P_1} 11|x_{Q_2}^2 11\} + \{x_{P_1}^2 11|x_{Q_2}^2 11\}\right] \quad (11)$$

The expressions of $\{x_{P_1}^2 11|x_{Q_2}^2 11\}$, $\{x_{P_1} 11|x_{Q_2}^2 11\}$, and $\{111|x_{Q_2}^2 11\}$ are respectively:

$$\{x_{P_1}^2 11|x_{Q_2}^2 11\} = \frac{\pi^{5/2}}{2\beta(\gamma_1 + \gamma_2)^{7/2}} \times \quad (12a)$$

$$\left\{4\beta^2 F_4(t) \overline{PQ}_x^4 - 12\beta F_3(t) \overline{PQ}_x^3 + \left[2(\gamma_1 + \gamma_2) \overline{PQ}_x^2 + 3\right] F_2(t) - \frac{(\gamma_1 + \gamma_2)}{\beta} F_1(t) + \frac{(\gamma_1 + \gamma_2)}{\beta} F_0(t)\right\} \\ \{x_{P_1} 11|x_{Q_2}^2 11\} = \frac{\pi^{5/2}}{\beta(\gamma_1 + \gamma_2)^{7/2}} \left\{2\gamma_2 \beta F_3(t) \overline{PQ}_x^3 - [2\gamma_2 F_2(t) + (\gamma_1 + \gamma_2) F_1(t)] \overline{PQ}_x\right\} \quad (12b)$$

$$\{111|x_{Q_2}^2 11\} = \frac{\pi^{5/2}}{\beta(\gamma_1 + \gamma_2)^{7/2}} \left\{2\gamma_2^2 F_2(t) \overline{PQ}_x^2 - \frac{\gamma_2(\gamma_1 + \gamma_2)}{\gamma_1} F_1(t) + \frac{(\gamma_1 + \gamma_2)^2}{\gamma_1} F_0(t)\right\}, \quad (12c)$$

where β is defined as $\frac{\gamma_1 \gamma_2}{(\gamma_1 + \gamma_2)}$ and the terms $F_n(t)$ are the Boys function:

$$F_n(t) = \int_0^1 x^{2n} \exp(-tx^2) dx \quad (13)$$

The evaluation of the Boys function had a recent renewed interest and was the subject of recent publications [12,13]. A different algorithm was developed for this work and will be discussed in a forthcoming paper.

The integrals $\left\{x_{P_1}^i y_{P_1}^j z_{P_1}^k | x_{Q_2}^i y_{Q_2}^j z_{Q_2}^k\right\}$ have important characteristics that can be explored to simplify the numerical approximations. The factors γ_1, γ_2 and β depend on the orbital exponents and are unaffected by geometrical changes. The Boys functions $F_n(t)$ depend on the orbital exponents and the separation between points \vec{P} and \vec{Q} , being independent of

the spatial orientation of the system. The factor \overline{PQ}_x (also y and z), which is the x component of the vector \overline{PQ} , depends on the orientation of the system. The "f" functions also introduce terms that depend on the orientation of the system, $\overline{PA}_x \overline{PB}_x$ and $(\overline{PA}_x + \overline{PB}_x)$ (see Table I and Eq. 11). The algorithm developed for the calculation of ERIs is based on the multivariate numerical approximation of all functions contributing to the integrals on the desired interval. The terms contributing to Eq. 11 consisting of products of Eqs 12a-12c and their respective "f" terms from Table I have complex spatial dependencies resulting from the Boys functions, $\overline{PA}_x \overline{PB}_x$, $(\overline{PA}_x + \overline{PB}_x)$ and \overline{PQ}_x terms. The strategy used in this work consists in recasting the parcels making the total ERI in terms of simpler functions which are products of the rotationally dependent functions \overline{PQ}_x , $\overline{PA}_x \overline{PB}_x$ and $(\overline{PA}_x + \overline{PB}_x)$, designated as g_n^{rot} , and a rotationally invariant term, G_n . The index n is the exponent of \overline{PQ}_x . In addition to the g_n^{rot}/G_n terms, there is an additional

rotationally dependent term derived from $\overline{PA_x PB_x}$, $g_{\overline{PA_x PB_x}}^{rot}$. The corresponding rotationally invariant term is desinated as $G_{n\overline{PA_x PB_x}}$. Using the terms of $\overline{PQ_x^4}$ for illustration, the rotation-

ally dependent functions g_4^{rot} and the corresponding rotationally independent term G_4 are calculated as:

$$g_4^{rot}(\overline{PQ_x^4}) = \frac{\sum_{a,b,c} c_a c_b c_c^2 N_a N_b N_c^2 \exp\left(-\frac{\alpha_1 \alpha_2 (\overline{AB})^2}{\gamma_1}\right) \times \frac{4\pi^{5/2}}{\beta(\gamma_1 + \gamma_2)^{7/2}} \times (\overline{PQ_x^4}) \times F_4(t)}{\sum_{a,b,c} c_a c_b c_c^2 N_a N_b N_c^2 \exp\left(-\frac{\alpha_1 \alpha_2 (\overline{AB})^2}{\gamma_1}\right) \times \frac{4\pi^{5/2}}{\beta(\gamma_1 + \gamma_2)^{7/2}} \times F_4(t)} \quad (14)$$

$$G_4 = \sum_{a,b,c} c_a c_b c_c^2 N_a N_b N_c^2 \exp\left(-\frac{\alpha_1 \alpha_2 (\overline{AB})^2}{\gamma_1}\right) \times \frac{4\pi^{5/2}}{\beta(\gamma_1 + \gamma_2)^{7/2}} \times F_4(t) \quad (15)$$

The important rotationally invariant term G_0 , which makes a direct contribution to the total computed ERI, is

$$G_0 = \frac{\pi^{5/2}}{\beta(\gamma_1 + \gamma_2)^{7/2}} \times \left[6F_2(t) - \frac{2(\gamma_1 + \gamma_2)}{\beta} F_1(t) + \frac{2(\gamma_1 + \gamma_2)}{\beta} F_0(t) - \frac{\gamma_2(\gamma_1 + \gamma_2)}{\gamma_1} F_1(t) + \frac{(\gamma_1 + \gamma_2)^2}{\gamma_1} F_0(t) \right] \quad (16)$$

III. MATHEMATICAL BACKGROUND

A. Multivariate approximation

In many applications, it is convenient to introduce *approximate functions*. An approximate function $g(x)$ is a function that given m data points x approximates the target values produced by the function $f(x)$ as closely as possible according to some metric. The approximant $g(x)$ is desired to be as smooth and compact as possible. The need to approximate often occurs when it is too costly or complex to use the true function, or even when the true function is unknown. The mathematical theory of approximation is well documented (see for example Ref. [14]). This work explores the possibility of approximating the complex and computationally costly Eq. 11 with simpler, and faster to evaluate, functions. All approximants are based on polynomial expansions (Eq. 17), in which the coefficients a_i are scalars and the generic basis functions $H(x)$ can take different forms:

$$f(x) = a_0 + a_1 H_1(x) + \dots + a_n H_n(x) \quad (17)$$

The main criterion to determine the quality of an approximation is the measurement of the “distance” between the target data points and the same set of points as obtained by the specified approximating function (approximant). It is important that the target and approximated points are as close as possible. A suitable metric to account for the global different between the set of true values and their respective approximations used extensively in this work is the l_2 -norm.

The multivariate approximation scheme developed to approximate ERIs consists of multiple levels of bivariate (or

univariate) approximants, with the fitting variables of a given level being expressed in terms of the variables of the next immediate level. The methodology is illustrated with the help of a 3-dimensional model depicted in Figure 1. To approximate the point $f(x_1, x_2, x_3)$, represented by the red sphere, a numerical approximant $g(x_1, x_2)$ of all points on the (x_1, x_2) plane is first developed. The function $g(x)$ is expanded in terms of primitive functions $H_n(x_1, x_2)$ according with Eq. 17. The dependency of x_3 , which is illustrated in Figure 1 by the vector originating at the blue sphere, is carried by fitting parameters a_i as functions of x_3 . In mathematical terms, the dependency of the fitting parameters a_i is given by another expansion similar to Eq. 17. The basis functions are represented by $H'_n(x_3)$ and the expansion has adjustable coefficients a_i :

$$a(x_3) = a'_0 + \sum_{i=1}^{n'} a'_i H'_i(x_3) \quad (18)$$

The process can be repeated multiple times, generating complex dependencies of multivariate functions. However, as the number of fitting parameters grow very fast with each additional layer of variable dependencies, in practice, the process is limited to a small number of layers.

In this work, the need for smooth functions arises because of the multiple dependencies of the variables. Since the fitting parameters carry additional dependencies themselves it is important that they are as smooth as possible to avoid discontinuities that make the next level fittings more complex. Other important criteria in defining the fitting process are computational efficiency, simplicity of algorithm implementation and future evolution of the method. When designing algorithms

Table I. Possible values of the f function as a function of l_1 , l_2 and the generic parameters A and B .

Eq. 9		
	$A \neq 0, B \neq 0$	$A = 0, B = 0$
$l_1 = 2, l_2 = 2$		
$i = 0$	$A^2 B^2$	0
$i = 1$	$2AB^2 + 2A^2 B$	0
$i = 2$	$B^2 + 4AB + A^2$	0
$i = 3$	$2B + 2A$	0
$i = 4$	1	1
$l_1 = 2, l_2 = 1$		
$i = 0$	$A^2 B$	0
$i = 1$	$2AB + A^2$	0
$i = 2$	$B + 2A$	0
$i = 3$	1	1
$l_1 = 2, l_2 = 0$		
$i = 0$	A^2	0
$i = 1$	$2A$	0
$i = 2$	1	1
$l_1 = 1, l_2 = 1$		
$i = 0$	AB	0
$i = 1$	$B + A$	0
$i = 2$	1	1
$l_1 = 1, l_2 = 0$		
$i = 0$	A	0
$i = 1$	1	1
$l_1 = 0, l_2 = 0$		
$i = 0$	1	1

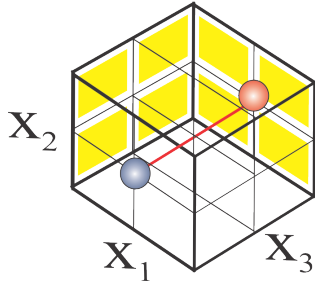


Figure 1. Illustration of the spatial dependency of multilayered approximating functions.

for numerical approximation is important to consider how fast and accurate the method is in the present, and to have a clear plan for future development.

The fitting functions were chosen to be bivariate Chebyshev orthogonal polynomials. Chebyshev polynomials form an important class of functions in curve fitting [15]. A similar expansion can be developed for surfaces $f(x, y)$ where the polynomial is based on to Chebyshev series with $\bar{x}, \bar{y} \in$

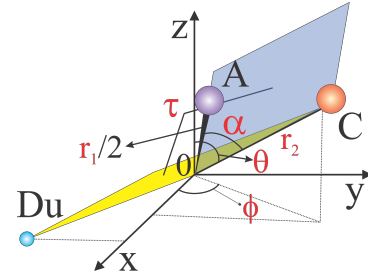


Figure 2. Illustration of the coordinates used in the fitting of three-center electron repulsion integrals

$$[-1, 1] \times [-1, 1]:$$

$$f(x, y) = \sum_{i=0}^{n_x} \sum_{j=0}^{n_y} a_{ij} T_i(\bar{x}) T_j(\bar{y}) \quad (19)$$

The arguments \bar{x} and \bar{y} are obtained from the original variables x and y by the transformations

$$\bar{x} = \frac{2x - (x_{\max} + x_{\min})}{x_{\max} - x_{\min}} \quad (20)$$

and

$$\bar{y} = \frac{2y - (y_{\max} + y_{\min})}{y_{\max} - y_{\min}} \quad (21)$$

The two-dimensional Chebyshev expansion was evaluated directly by computing the polynomials and summing all contributions according to Eq. 19.

B. Choice of coordinates

Each of the terms g_n^{rot} and G_n required to calculate ERIs according to the prescription of Eqs 14-16 can be expressed in terms of a finite number of variables. Fitting of three-center ERIs requires six coordinates that are used to position the atomic centers carrying the basis functions. Importantly, only the total number of variables has to be fulfilled and not the nature of the individual coordinates as long as they provide the spatial assignment of the atomic centers. It is, however, advisable to use combinations of variables that lead to simpler fitting expressions, in addition to having physical meanings that can be related to common geometrical transformations. In this respect bond distances, angles and torsions are prime candidates.

The protocol followed in this work consists in separating the rotationally dependent terms of PQ_x^n and $PA_x PB_x$ from the rotationally invariant counterparts. The set of coordinates chosen for the fitting of the rotationally invariant part are two distances, r_1 and r_2 , and the internal angle α . r_1 is the distance between the centers A and B and r_2 is the separation between C and the midpoint of \vec{AB} represented by \vec{O} . α is the angle $\hat{C}\hat{O}\hat{A}$

(see Figure 2). The projections \overline{PQ}_x , \overline{PQ}_y , or \overline{PQ}_z require special attention since they impart the rotational invariance of the integrals. Their spatial dependencies are significantly more complex, requiring an extra set of coordinates. The extra variables are the polar spherical coordinates θ and φ , which are used to position the atomic center C and the dihedral angle τ , which is used to determine the relative orientation of centers A (and B) relative to C (see Figure 2).

IV. RESULTS AND DISCUSSION

The following Section is dedicated to evaluating the accuracy and speed of the numerical algorithm to approximate ERIs. Emphasis is placed on testing the ability of the method to accurately reproduce the integral $(p_{xA}p_{xB}|p_{xC}p_{xC})$ with carbon atoms are placed at A, B, and C positions. All calculations were based on the STO-6G basis set. This basis set is sufficiently small to allow computation of the many target ERIs used in the parameterization in a reasonable time. All calculations were done on modest hardware: AMD 8350 CPU and 24 GByte of RAM memory. No parallelization was attempted and the calculations were done on a single-core. All codes were compiled with Gfortran using the `-O3` compiler flag. The approximation of the rotationally independent terms is discussed first, with G_4 being used as example. Afterward, the fitting of the rotationally dependent terms is analyzed. The approximating methodologies are illustrated with the help of $g_4^{rot}(\overline{PQ}_x^4)$ since it is representative of the other terms. The accuracy and speed of the multivariate methodology of approximation are discussed in Sect. IV B and IV C. Two quantities are used to measure the goodness-of-fit of the approximants: the Root Mean Square Error (RMSE) and R^2 . RMSE measures the total deviation of the computed from the target values, and a value closer to zero indicates the fit is good and is useful for prediction. Another quantity to access the quality of an approximation is R^2 , which indicates how well the approximation explains variation in the data. The closest the value of R^2 is to one the better is the approximation. The domains of the variables influencing the rotation of the systems are $\alpha \in [0, 180^\circ]$, and $\tau, \theta, \varphi \in [0, 90^\circ]$. The domains of τ, θ , and φ are limited to 90° because of the symmetry relations resulting from having the same element at the positions A, B, and C. Figure S1 illustrates the symmetry effects for the dependencies of (α, τ) and (θ, φ) for the function $g_4^{rot}(\overline{PQ}_x^4)$.

A. Fitting the rotationally independent terms $G_n(\alpha, r_1, r_2)$ and $G_{\overline{PA}_x\overline{PB}_x}(\alpha, r_1, r_2)$

The protocol described in Sect. III A for the multivariate fitting of the different parcels making the analytical expression of the ERIs starts with the initial fitting of rotationally invariant functions $G_n(\alpha, r_1, r_2)$ and $G_{\overline{PA}_x\overline{PB}_x}(\alpha, r_1, r_2)$. In most cases, these are auxiliary functions used to create smoother rotationally dependent surfaces that are easier to fit, although $G_0(\alpha, r_1, r_2)$ contributes directly to the final integral

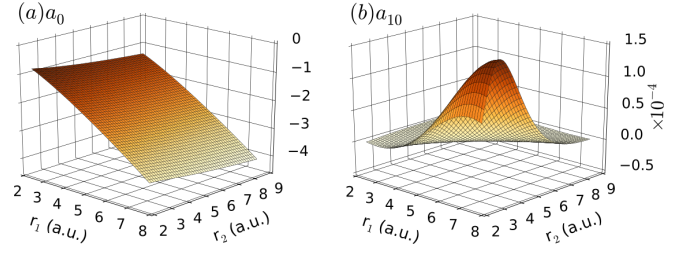


Figure 3. Illustration of the spatial dependency of a_0 , and a_{10} of the rotationally independent term $G_4(\alpha, r_1, r_2)$. The surfaces are smooth and suitable for accurate approximation. The units of the surfaces are radians^{-1} .

(Eq. 16). The dependencies of G_n are on the angle α and distances r_1 and r_2 . The protocol for the fitting of $G_n(\alpha, r_1, r_2)$ and $G_{\overline{PA}_x\overline{PB}_x}(\alpha, r_1, r_2)$ calls to the initial fitting of the α dependency. The plots of the $G_n(\alpha, r_1, r_2)$ functions relative to α (with r_1 and r_2 fixed) follows a similar symmetric sinusoidal curve. The dependency of $G_n(\alpha, r_1 = 2.6 \text{ a.u.}, r_2 = 5.0 \text{ a.u.})$ is illustrated on Figure S2. The function of choice for the fitting of the dependency of α , in radians, was a 10th-order polynomial written in the form:

$$f(\alpha) = a_0(r_1, r_2) + a_1(r_1, r_2)\alpha + \dots + a_{10}(r_1, r_2)\alpha^{10} \quad (22)$$

The dependency of each a_i term on (r_1, r_2) is highlighted in Eq. 22, in accordance with the multivariate fitting algorithm described in Sect. III A. The 10th order expansion was found to be acceptable in terms of computational cost and accuracy. Because of the dependence of the coefficients on the distances r_1 and r_2 , it is important to keep the polynomial expansion above as compact as possible.

The dependency of the coefficients a_0 and a_{10} in Eq. 22 on (r_1, r_2) is illustrated in Figure 3. The key assumption of this work, that the fitting coefficients of a polynomial approximant at a certain level have smooth spatial dependencies of the variables of the next level, and thus, are able to carry that spatial dependency, is fully fulfilled. Although no rigorous mathematical proof is presented, the surfaces of a_0 , and a_{10} , as well as the surfaces of the remaining coefficients, are smooth and can be approximated using the multivariate techniques presented before. Each coefficient $a_i(r_1, r_2)$ was fitted by double Chebyshev polynomials with arguments \bar{r}_1 and \bar{r}_2 (see Eq. 19). The order of the expansion was truncated at 14. It is the largest Chebyshev polynomial order used in this work. The computational cost of using such a long expansion is not prohibitive for two reasons. First, each increase of the Chebyshev polynomial order only contributes the additional number of parameters times the order of the univariate polynomial expansion of α . Second, the rotationally invariant terms only need to be calculated once and stored. The same rotationally invariant terms, $G_n(\alpha, r_1, r_2)$ and $G_{\overline{PA}_x\overline{PB}_x}(\alpha, r_1, r_2)$, can be used in the fitting of all ERIs whether they involve p_x , p_y or p_z functions. The explicit expression for fitting all spatial dependencies of $G_n(\alpha, r_1, r_2)$, combining Eq. 22 above with the

bivariate (r_1, r_2) Chebyshev polynomial for each of the coefficients a_i is

$$G_n(\alpha, r_1, r_2) / G_{\overline{PA_x} \overline{PB_x}}(\alpha, r_1, r_2) = \left(\sum_{i=0}^{n_x} \sum_{j=0}^{n_y} a_{ij}^{(0)} T_i(\bar{r}_1) T_j(\bar{r}_2) \right) + \left(\sum_{i=0}^{n_x} \sum_{j=0}^{n_y} a_{ij}^{(1)} T_i(\bar{r}_1) T_j(\bar{r}_2) \right) \alpha + \dots + \left(\sum_{i=0}^{n_x} \sum_{j=0}^{n_y} a_{ij}^{(9)} T_i(\bar{r}_1) T_j(\bar{r}_2) \right) \alpha^9 + \left(\sum_{i=0}^{n_x} \sum_{j=0}^{n_y} a_{ij}^{(10)} T_i(\bar{r}_1) T_j(\bar{r}_2) \right) \alpha^{10} \quad (23)$$

where $T_i(\bar{r}_1)$ is the Chebyshev polynomial of the first kind of degree i with argument \bar{r}_1 , and $T_j(\bar{r}_2)$ is similarly defined for j and \bar{r}_2 . The approximations are extremely accurate with overall RMSEs lower than 7.35E-05 and R^2 coefficients $\gg 0.99999$. The residuals for $G_n(\alpha, r_1, r_2)$ and $G_{\overline{PA_x} \overline{PB_x}}(\alpha, r_1, r_2)$ are plotted in Figure S3 for $\alpha = 145^\circ$.

B. Fitting the rotationally dependent terms $g_n^{\text{rot}}(\overline{PQ_x^n})$ and $g_{\overline{PA_x} \overline{PB_x}}^{\text{rot}}$

The rotationally dependent functions hold the effects of the $\overline{PQ_x^n}$, $\overline{PA_x} \overline{PB_x}$ and $(\overline{PA_x} + \overline{PB_x})$ terms in the three-center ERI of the kind $(p_{xA} p_{xB} | p_{xC} p_{xC})$. These are considerably more challenging to approximate since they depend on six variables instead of the three variables in the rotationally independent terms. Two quantities were defined to measure the contribution of each term to the total ERI: the Average Absolute Percentage Contribution (AAPC) and Maximum Absolute Percentage Contribution (MAPC). The absolute value of each term was chosen because each can be positive or negative. The AAPC and MAPC quantities are calculated for

a generic term f_a as respectively $100 \times \sum_i^n \frac{f_{a,i}}{(f_{1,i} + \dots + f_{k,i})^n}$ and $100 \times \max \frac{f_{a,i}}{(f_{1,i} + \dots + f_{k,i})^n}$. Values closer to 100% indicate a stronger contribution to the integral and likewise values close to zero mean smaller contributions. The most significant contributors are the rotationally independent term G_0 and the rotationally dependent function of $\overline{PA_x} \overline{PB_x}$. Interestingly, all the terms containing the projections $\overline{PQ_x}$ make considerably smaller contributions (see Table II). According to the relevance of each term, different expansions can be defined without imparting significantly the accuracy of the final approximation. The G_0 term is already fitted with the highest order of any Chebyshev polynomial, and the accuracy of the approximations can be hardly improved.

The fitting protocol for $g_n^{\text{rot}}(\overline{PQ_x^n})$ and $g_{\overline{PA_x} \overline{PB_x}}^{\text{rot}}$ requires three layers of fittings using bivariate Chebyshev polynomials. The pairing of variables is: $(\theta, \varphi) \rightarrow (\alpha, \tau) \rightarrow (r_1, r_2)$. In the first step, fitting functions $f(\theta, \varphi)$ are determined for each of the target points (α, τ, r_1, r_2) (see Eq. 24a). The coefficients a_{ij} carry the dependency of the remaining variables α , τ , r_1 , and r_2 . In the second level of optimization, each of the coefficients a_{ij} is fitted similarly with bivariate Chebyshev polynomial (Eq. 24b). The fitting coefficients b_{kl}^{ij} carry the dependency of (r_1, r_2) and each will be fitted in the third level of fittings (Eq. 24c).

$$g^{\text{rot}} \simeq f(\theta, \varphi)_{|\alpha^0, \tau^0, r_1^0, r_2^0} = \sum_{i=0}^{n_\theta} \sum_{j=0}^{n_\varphi} a_{ij}(\alpha, \tau, r_1, r_2) T_i(\bar{\theta}) T_j(\bar{\varphi}) \quad (24a)$$

$$a_{ij}(\alpha, \tau)_{|r_1^0, r_2^0} = \sum_{k=0}^{n_\alpha} \sum_{l=0}^{n_\tau} b_{kl}^{ij}(r_1, r_2) T_k(\bar{\alpha}) T_l(\bar{\tau}) \quad (24b)$$

$$b_{kl}^{ij}(r_1, r_2) = \sum_{m=0}^{n_{r_1}} \sum_{n=0}^{n_{r_2}} c_{mn}^{ij,kl} T_m(\bar{r}_1) T_n(\bar{r}_2) \quad (24c)$$

In Eq. 24 the superscript “0” means that the corresponding variable assumes a fixed value.

Figure 4 illustrates selected surfaces $f(\theta, \varphi)$ for specific values of (r_1, r_2) and (α, τ) . All surfaces have similar

Table II. Average Absolute Percentage Contribution (AAPC) and Maximum Absolute Percentage Contribution (AAPC) of each term contributing to $(p_{xA}p_{xB}|p_{xC}p_{xC})$

	Term						
	\overline{PQ}_x^4	\overline{PQ}_x^3	$\overline{PQ}_x^2(l_1+l_2=0)$	\overline{PQ}_x	$\overline{PQ}_x^2(l_1+l_2=2)$	$\overline{PA}_x\overline{PB}_x$	G_0
AAPC	0.4	0.1	0.8	0.4	1.0	45.7	51.6
MAPC	12.7	2.2	10.1	2.5	11.0	91.2	100.0

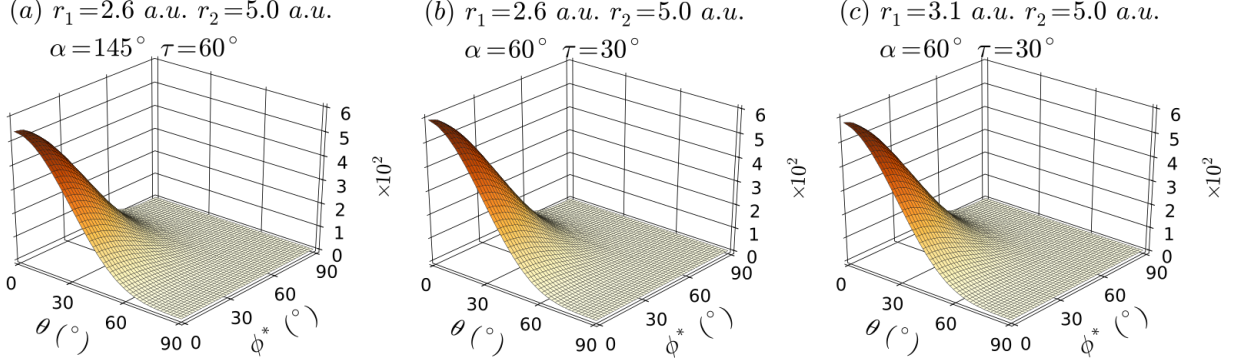


Figure 4. Illustration of the dependency of θ and ϕ for fixed values of r_1 , r_2 , α , and τ for $g_4^{rot}(\overline{PQ}_x^4)$. (a) and (b) show the effect of varying α and τ , whereas (b) and (c) illustrate the effect of varying r_1 , r_2 . Approximation of each surface will require calculation of the coefficients a_{ij} of Equation 24a.

Gaussian-like shapes with maxima at $(0^\circ, 0^\circ)$. It is noteworthy that to facilitate the numerical approximations, the surfaces were symmetrized through the change of coordinate $\phi^* = 180^\circ - \phi$. The functions of \overline{PQ}_x^n were approximated with Chebyshev polynomials of order 10. The rotationally dependent function of $\overline{PA}_x\overline{PB}_x$ was approximated with Chebyshev polynomials of order 10 and 11 and the overall results are discussed in Sect. IV C.

Despite the apparent similarity of the $f(\theta, \phi)$ functions for fixed values of (r_1, r_2) the coefficients a_{ij} show remarkable variability as a function of α and τ . Figure 5 illustrates the coefficients a_1 and a_{66} for $r_1 = 2.6$ a.u. and $r_2 = 5.0$ a.u.. Similarly to the approximation of the rotationally independent terms, the bivariate surfaces $a_{ij}(\alpha, \tau)$ are smooth and, thus, easily approximated by bivariate Chebyshev polynomials. The operational parameters for the approximation of the $f(\theta, \phi)$ and $a_{ij}(\alpha, \tau)$ surfaces are given in Table III. It was opted to approximate the $a_{ij}(\alpha, \tau)$ and $f(\theta, \phi)$ surfaces with the Chebyshev polynomials of the same order.

The final step in the approximation of the rotationally dependent functions is the fitting of the coefficients b_{kl}^{ij} . In Figure 6, the surfaces corresponding to b_1^1 and b_1^{66} for $g_4^{rot}(\overline{PQ}_x^4)$ are shown as a function of the remaining coordinates r_1 and r_2 . The dependency of the b_{kl}^{ij} coefficients is considerably simpler with the function being monotonically increasing in r_2 (i.e. for fixed values of r_1). Two important simplifications can be introduced in the approximation of the b_{kl}^{ij} . First, the order of the Chebyshev polynomials can be reduced since the surfaces are easy to approximate, and second, many of the coefficients can be eliminated since their contribution to $a_{ij}(\alpha, \tau)$ is negligible. Three parameters ϵ_i are used to define

three regions of different approximating accuracy. In practice, the maximum absolute value of $b_{kl}^{ij}(r_1, r_2)$ is compared with each parameter ϵ_i . If $\max |b_{kl}^{ij}(r_1, r_2)| \geq \epsilon_1$ the surface is approximated with the “fine” expansion of Chebyshev polynomials. “Medium” and “coarse” expansions are used when $\epsilon_1 > \max |b_{kl}^{ij}(r_1, r_2)| \geq \epsilon_2$ and $\epsilon_2 > \max |b_{kl}^{ij}(r_1, r_2)| \geq \epsilon_3$. Surfaces for which $\max |b_{kl}^{ij}(r_1, r_2)| < \epsilon_3$ are discarded and the corresponding coefficients $c_{mn}^{ij,kl}$ are zero. The values of the ϵ_i parameters and the orders of the Chebyshev expansions for the “fine”, “medium” and “coarse” regions are also in Table III. The quality of the different approximants of $f(\theta, \phi)$, $a_{ij}(\alpha, \tau)$, and $b_{kl}^{ij}(r_1, r_2)$ is gauged in Table IV, where goodness-of-fit results are presented and compared for the g_n^{rot} term. A distinction was made between the $g_n^{rot}(\overline{PQ}_x^n)$ and $g_n^{rot}(\overline{PA}_x\overline{PB}_x)$ terms because of their very different contributions to the total computed ERI. Because of the small contributions to the computed ERI the \overline{PQ}_x^n terms make, the long expansions of order 10 that were used in the approximation of the $f(\theta, \phi)$ and $a_{ij}(\alpha, \tau)$ surfaces were probably overkill. The RMSE and R^2 values are nevertheless excellent, indicating overall accurate approximations. Special care was placed on the approximation of $g_n^{rot}(\overline{PA}_x\overline{PB}_x)$ because of the very significant partial contribution to the total ERI. Thus, the rotational invariance of the approximated ERI greatly depends on $g_n^{rot}(\overline{PA}_x\overline{PB}_x)$. From Table IV, the most stringent model 5 seems necessary to achieve an excellent accuracy in the approximation. It is important to stretch that approximation with bivariate Chebyshev polynomials provides a way to systematically improve the quality of the approximation. For example, the RMSE decreased an order of magnitude on going from

Table III. Operational parameters for the different levels of approximation tested

Functions of $g_n^{rot}(\overline{PQ}_x^n)$							
Model	Order*	Threshold			Order of Chebyshev polynomial for b_{kl}^{ij}		
		ϵ_1	ϵ_2	ϵ_3	Fine	Medium	Coarse
1	10	1.0	1.0E-02	1.0E-06	8	6	4
2	10	1.0E-01	1.0E-03	1.0E-07	8	6	4
Function of $g_n^{rot}(\overline{PA}_x\overline{PB}_x)$							
		Threshold			Order of Chebyshev polynomial for b_{kl}^{ij}		
		ϵ_1	ϵ_2	ϵ_3	Fine	Medium	Coarse
3	10	1.0E-01	1.0E-03	1.0E-07	8	6	4
4	11	1.0E-01	1.0E-03	1.0E-07	8	6	4
5	11	1.0E-03	1.0E-06	1.0E-09	10	8	6
*Order of Chebyshev polynomial for $f(\theta, \varphi)$ and $a_{ij}(\alpha, \tau)$							

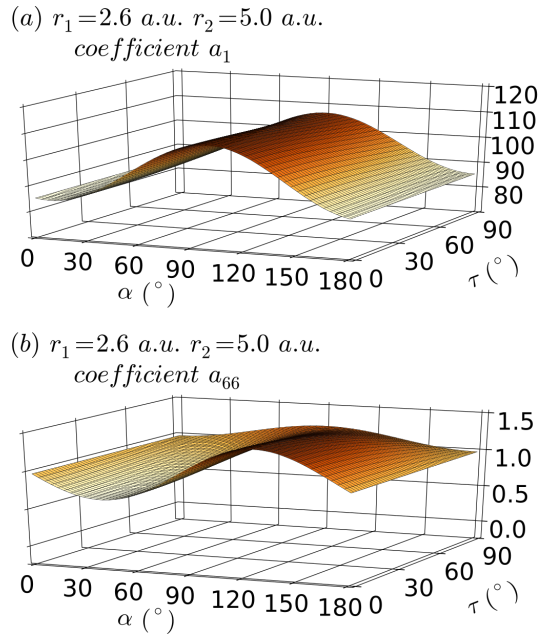


Figure 5. Illustration of dependency of the coefficients $a_{ij}(\alpha, \tau)$ of Equation 24a on the angles α and τ for fixed values of $r_1 = 2.60$ a.u. and $r_2 = 5.0$ a.u. for $g_4^{rot}(\overline{PQ}_x^4)$. The surfaces are smooth and have distinctive magnitudes that will be explored to reduce the order of the polynomials used in the fittings.

models 3 and 4 to model 5.

C. Adding all together: assembly of the computed $(\mathbf{p}_{xA}\mathbf{p}_{xB}|\mathbf{p}_{xC}\mathbf{p}_{xC})$ two-electron-repulsion integral

The culmination of this work on the numerical approximation of ERIs is the assembly of the calculated values of the $(\mathbf{p}_{xA}\mathbf{p}_{xB}|\mathbf{p}_{xC}\mathbf{p}_{xC})$ ERIs from the different terms discussed above and the comparison with the real analytical values. The contributors to the calculated ERI are: the rotationally invariant term G_0 , the rotationally dependent terms $g_n^{rot}(\overline{PQ}_x^n) \cdot G_n$

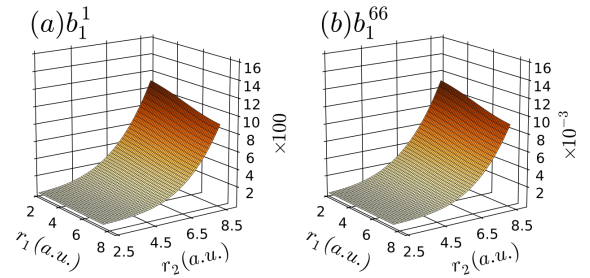


Figure 6. Illustration of the dependency of the coefficients $b_{kl}^{ij}(r_1, r_2)$ of Equation 24b on the distances r_1 and r_2 for $g_4^{rot}(\overline{PQ}_x^4)$. The surface on the left (a) is for $b_1^1(r_1, r_2)$ and the surface on the right (b) is for $b_1^{66}(r_1, r_2)$. Both surfaces are monotonically increasing in the r_2 direction, i.e. for fixed r_1 . These surfaces can be approximated with more compact Chebyshev polynomials, thus reducing the overall computational cost.

and $g_n^{rot}(\overline{PA}_x\overline{PB}_x) \cdot G_{\overline{PA}_x\overline{PB}_x}$. The approximation of G_0 was unique, using the highest order expansion of this work. For the terms derived from $g_n^{rot}(\overline{PQ}_x^n)$, two models were analyzed, but only model 2 was included in the assembly of the final approximated ERI. The three models tested in the approximation of $g_n^{rot}(\overline{PA}_x\overline{PB}_x)$ were included in the computation of the approximated ERI. Goodness-of-fit results for the total ERI are shown in Table V. The results are the expected, and follow the pattern obtained for $g_n^{rot}(\overline{PA}_x\overline{PB}_x)$ (Table IV). It is important to stretch not only the magnitude, but also the distribution of the residuals. In Figure 7, the residuals for $r_1 = 2.6$ a.u., $r_2 = 5.0$ a.u., $\alpha = 60/145^\circ$ and $\tau = 30/60^\circ$ are plotted as a function of θ and φ for $(G_0 + \text{model 2} + \text{model 4})$ and $(G_0 + \text{model 2} + \text{model 5})$. The biggest residual is roughly one order of magnitude smaller with model 5 than with model 4, but importantly with model 5 the residuals are less than $1.0\text{E-}04$ for most of the θ and φ domains, being higher in a very restricted area around $\theta = 0^\circ$ and $\varphi = 90^\circ$. In contrast, with the less accurate model 4 (also with the similar model 3) the residuals have significantly higher values across the whole domains of θ and φ . The same pattern was found for other points and

Table IV. Goodness-of-fit estimates and CPU times for terms $g_n^{rot}(\overline{PQ}_x^n)$, $n = 1, 2, 3, 4$ and $g_n^{rot}(\overline{PA}_x\overline{PB}_x)$

$g_n^{rot}(\overline{PQ}_x^n)$, $n = 1, 2, 3, 4$					
	\overline{PQ}_x^4	\overline{PQ}_x^3	$\overline{PQ}_x^2(l_1 + l_2 = 0)$	\overline{PQ}_x	$\overline{PQ}_x^2(l_1 + l_2 = 2)$
Model 1					
RMSE	4.03E-06	7.81E-06	2.82E-05	3.28E-05	1.46E-05
R ²	0.99997	0.99821	0.99950	0.97363	0.99973
time (s)	555.7±3.4	546.4±3.9	550.0±3.3	549.4±2.4	554.4±4.3
Model 2					
RMSE	7.94E-07	1.18E-06	4.63E-06	4.99E-06	3.29E-06
R ²	> 0.99999	0.99996	0.99999	0.99943	0.99999
time (s)	557.3±6.9	549.4±2.4	555.0±4.5	562.3±3.7	564.1±7.8
			$g_n^{rot}(\overline{PA}_x\overline{PB}_x)$		
			RMSE	R ²	time (s)
Model 3			6.07E-04	0.99929	557.4±6.2
Model 4			6.06E-04	0.99929	552.9±4.3
Model 5			1.59E-05	>> 0.99999	706.2±12.6

Table V. Goodness-of-fit estimates for the approximated $(p_{xA}p_{xB}|p_{xC}p_{xC})$ ERI

	RMSE	R ²
G_0 + Model 2 + Model 3	6,07E-04	0,99929
G_0 + Model 2 + Model 4	6,06E-04	0,99929
G_0 + Model 2 + Model 5	1,59E-05	>> 0,99999

in the future an exhaustive statistical study will be performed to determine the validity of these anecdotal observations. The implications are tremendous since it suggests the possibility of extending the areas of extreme accuracy by redefining the limits of the domains where the Chebyshev polynomials are defined.

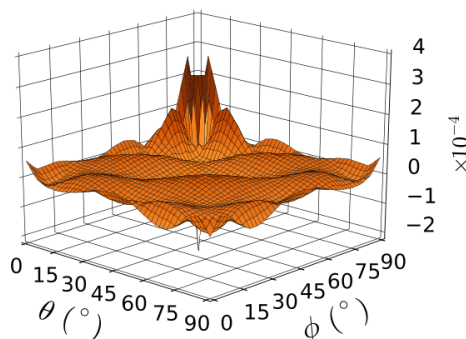
The timings of the calculation of the different rotationally dependent terms are given in Table IV. Despite the amazing results, with speedups of 4-5 orders of magnitude, consideration of the timings required for the numerical approximation of ERIs is secondary in this work. The calculations were performed taking advantage of the highly structured grid points and are hardly representative of real world scenarios. However, no specific optimizations were developed and the Chebyshev polynomials were performed directly using Eq. 19. In the future, specific optimizations will be introduced. For example, the most costly calculation of the b_{kl}^{ij} terms according to Eq. 24c can be vectorized, introducing considerable speed gains. The newly developed algorithm is inherently fast, requiring only matrix-vector or matrix-matrix multiplications, operations that are highly optimized on multiple computer architectures, including GPUs.

V. CONCLUSIONS AND FUTURE PROSPECTS

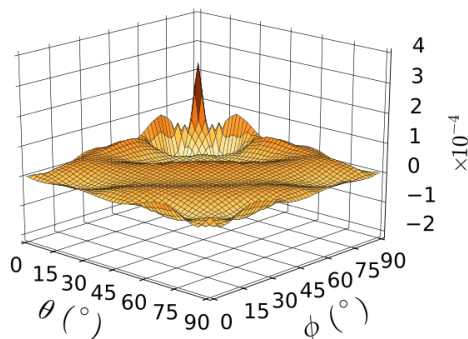
The work presented in this publication is the first step of a large effort to develop novel tight binding computational methodologies that are able to study large, complex systems. In the path to faster and more generic computational quantum methods, three aspects are the most significant: 1) computation of ERIs, theoretically an $O(N^4)$ process, 2) diagonalization, itself an $O(N^3)$ process, and 3) the SCF iterations. The focus of this work was on the efficient and accurate computation of ERIs. The approach consisted in using multivariate approximation techniques to reproduce pre-computed target ERI data. It is a proof-of-concept work aimed at demonstrating the feasibility of such approximations. To my best knowledge, this was the first time that such techniques were published. The test system was the three-center ERI $(p_{xA}p_{xB}|p_{xC}p_{xC})$ with all atoms being carbons. Having all atoms the same, introduces important symmetry relations that help to simplify the amount of data required for the approximations. In the initial phase of development, when multiple calculations were needed in order to generate adequate target data, it was important to keep the number of calculations to a minimum. The same holds for the target basis set. The small STO-6G was used because it allows efficient calculation of the many ERIs required as target data. The methodology for the numerical approximation consisted in decomposing a six-variable problem and a three-variable problem into three bivariate problems and one univariate plus one bivariate problem, respectively. The chosen approximating functions were bivariate Chebyshev polynomials and a univariate polynomial of order 10. The assumption was that for each sequential variable reduction, the approximating coefficients yield a continuous function that can be approximated by another set of polynomial approximants. The feasibility of the methodology relies on assuming that the approximating coefficients of a certain layer determine a continuous surface that can be fitted in the next layer of approx-

Model 2 + Model 5

(a) $r_1 = 2.6 \text{ a.u.}$ $r_2 = 5.0 \text{ a.u.}$
 $\alpha = 60^\circ$ $\tau = 30^\circ$

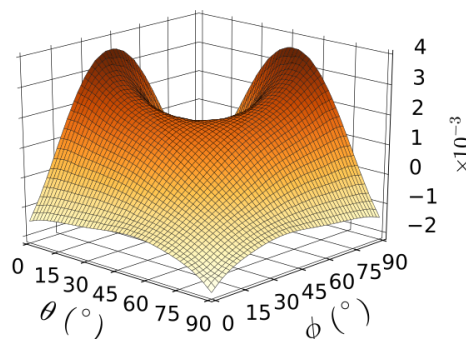


(b) $r_1 = 2.6 \text{ a.u.}$ $r_2 = 5.0 \text{ a.u.}$
 $\alpha = 145^\circ$ $\tau = 60^\circ$



Model 2 + Model 4

(c) $r_1 = 2.6 \text{ a.u.}$ $r_2 = 5.0 \text{ a.u.}$
 $\alpha = 60^\circ$ $\tau = 30^\circ$



(d) $r_1 = 2.6 \text{ a.u.}$ $r_2 = 5.0 \text{ a.u.}$
 $\alpha = 145^\circ$ $\tau = 60^\circ$

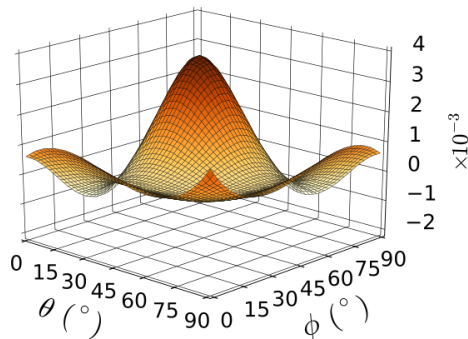


Figure 7. Illustration of the residuals of the total computed ERI for two different models of Table III: model 2 + model 4 and model 2 + model 5. The rotationally dependent terms $g_n^{rot}(\overline{PQ}_x^n)$ were only fitted with the parameters of model 2. In contrast, the term $g_n^{rot}(\overline{PA}_x\overline{PB}_x)$ was fitted with both model 4 and model 5. The impact of model 5 on the overall accuracy is impressive, with the residuals being one order of magnitude smaller than with model 4 and importantly the biggest absolute residuals very localized near $\theta = 0^\circ$ and $\phi = 90^\circ$

imations using the same technique. Although no mathematical justification was attempted, it was indeed verified that all surfaces and the single curve are continuous and could be easily approximated. It is important to remember that the novel methodology to approximate ERIs is not general and is not intended to replace existing basis sets.

The results are excellent with very small errors. In plots of residuals for two specific points, it was found that for most of the approximating domains the absolute error was significantly less than $1.0\text{E-}04$. This means that the approximating methodology is able to maintain the rotational invariance of the computed integrals. Importantly, the new approach does not depend on the size of the contractions of the basis set. Although it was not a priority of this work, and no special attempts were made to optimize the speed of the numerical approximations, the methodology is very fast. Normalizing the CPU time to 1 core of the computer system used in the work (AMD 8350), computation of the total number of ERIs used in the fittings, more than 937 million, could be completed in

minutes. The same calculation of the analytical ERIs on the same hardware would require approximately three months.

The first major development in the future will be the creation of a library of approximated ERIs, starting with the most common elements in Biology: H, C, N, O, S, Na, K, Cl, Fe, Zn, Cu, and Ni. Appropriated long single- and double- ζ basis sets will also need to be developed for these elements. A new tight-binding approach will be implemented. It will incorporate the new methodology for fast diagonalizations that was recently developed and will be linked to the library of approximated ERIs. In a future publication, the new efficient diagonalization techniques will be described.

In conclusion, this work opens new perspectives to the future of computational chemistry in general, for example to molecular simulations of large, complex systems. The efficient computation of ERIs eliminates a significant barrier to the generalization of computational quantum methods to large systems. The combination of the methods for fast evaluation of ERIs with novel approaches to diagonalize very large ma-

trices will allow development of specialized quantum based methodologies that will be simultaneously fast and accurate.

ACKNOWLEDGMENTS

P.E.M.L. wishes to thank M.M.G and J.D.N for support and M.S.L. for reading the manuscript.

-
- [1] Dirac PAM (1929) Quantum Mechanics of Many-Electron Systems. *Proc R Soc Lond A* 123: 714-733.
 - [2] Lopes PEM, Huang J, Shim J, Luo Y, Li H, et al. (2013) Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *J Chem Theory Comput* 9: 5430-5449.
 - [3] Thomson AJ, Gray HB (1998) Bio-inorganic chemistry. *Curr Opin Chem Biol* 2: 155-158.
 - [4] Boys SF (1950) Electronic Wave Functions. I. A General Method of Calculation for the Stationary States of Any Molecular System. *Proc R Soc Lond A* 200: 542.
 - [5] Dupuis M, Rys J, King HF (1976) Evaluation of molecular integrals over Gaussian basis functions. *J Chem Phys* 65: 111-116.
 - [6] King HF, Dupuis M (1976) Numerical integration using rys polynomials. *J Comput Phys* 21: 144-165.
 - [7] McMurchie LE, Davidson ER (1978) One- and two-electron integrals over cartesian gaussian functions. *J Comput Phys* 26: 218-231.
 - [8] Obara S, Saika A (1986) Efficient recursive computation of molecular integrals over Cartesian Gaussian functions. *J Chem Phys* 84: 3963-3974.
 - [9] Reine S, Helgaker T, Lindh R (2012) Multi-electron integrals. *WIREs Comput Mol Sci* 2: 290-303.
 - [10] Clementi E, Davis DR (1966) Electronic structure of large molecular systems. *J Comput Phys* 1: 223-244.
 - [11] Clementi E (1991) Modern techniques in computational chemistry : MOTECC-91. Leiden, the Netherlands: ESCOM.
 - [12] Guseinov II, Mamedov BA (2006) Evaluation of the Boys Function using Analytical Relations. *J Math Chem* 40: 179-183.
 - [13] Weiss AKH, Ochsenfeld C (2015) A rigorous and optimized strategy for the evaluation of the Boys function kernel in molecular electronic structure theory. *J Comput Chem* 36: 1390-1398.
 - [14] Powell MJD (1981) Approximation theory and methods. Cambridge [England]; New York: Cambridge University Press.
 - [15] Mason JC, Handscomb DC (2003) Chebyshev polynomials. Boca Raton, Fla.: Chapman & Hall/CRC.